Journal of Statistical Theory and Practice (2020) 14:26
https://doi.org/10.1007/s42519-020-00097-5

ORIGINAL ARTICLE

# Nonparametric Predictive Inference Bootstrap with Application to Reproducibility of the Two-Sample Kolmogorov–Smirnov Test

**Frank P. A. Coolen**[1] · **Sulafah Bin Himd**[2]



**Abstract**
This paper introduces a new bootstrap method based on the nonparametric predictive inference (NPI) approach to statistics. NPI is a frequentist statistics framework which explicitly focuses on prediction of future observations. The NPI framework enables a bootstrap method (NPI-B) to be introduced which, different to Efron's classical bootstrap (Ef-B), is aimed at prediction of future observations instead of estimation of population characteristics. A brief initial comparison of NPI-B and Ef-B is presented. The main reason for introducing NPI-B here is for its application to NPI for reproducibility of statistical tests, which is illustrated for the two-sample Kolmogorov–Smirnov test.

**Keywords** Bootstrap · Kolmogorov–Smirnov test · Nonparametric predictive inference · Reproducibility of tests

## 1 Introduction

Nonparametric predictive inference (NPI) [3, 8, 9] is a frequentist statistical methodology based on only few assumptions. For general inferences on one or more future real-valued observations, based on $n$ data observations, it uses Hill's assumption $A_{(n)}$ [17] in combination with imprecise probabilities to quantify uncertainty [4]. Augustin and Coolen [3] showed that NPI-based lower and upper probabilities have strong consistency properties in imprecise probability theory. They always contain the corresponding empirical probability, and the

✉ Frank P. A. Coolen
frank.coolen@durham.ac.uk

Sulafah Bin Himd
shamad@kau.edu.sa

[1] Department of Mathematical Sciences, Durham University, Durham, UK

[2] Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia

Published online: 28 April 2020

imprecision, defined as the difference between the upper and lower probabilities for an event, logically reflects the amount of data available. In this paper, we present an alternative to the classical bootstrap method [14, 15], based on NPI. The method is actually quite close in nature to Banks' smoothed bootstrap [5], but with the crucial difference that the values in one bootstrap sample are not derived conditionally independently, given the original data. While established bootstrap methods focus on estimation of characteristics of an assumed underlying population from which observations are randomly drawn, the new NPI bootstrap method, indicated by NPI-B, is explicitly aimed at predictive inference, with variability in different bootstrap samples reflecting uncertainty in prediction in line with the NPI method.

The bootstrap method was introduced by Efron [14]. It is a resampling technique for estimating characteristics of an assumed population from which the data observations were sampled, and for quantifying the quality of the estimates by providing an indication of the variability involved. The bootstrap method has become one of the most used statistical methods. It uses Monte Carlo sampling to generate an empirical estimate of the sampling distribution of the statistic of interest, the bootstrap distribution. It uses a plug-in principle to approximate the sampling distribution by the bootstrap distribution. Efron [14] defined a bootstrap sample $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$, obtained by randomly sampling $n$ times, with replacement, from the original data points $x_1, x_2, \ldots, x_n$.

There are many references that show the principles and validity of bootstrap and how it works. Efron and Tibshirani [15] and Davison and Hinkley [12] have described bootstrap methods with example applications to statistical tests, confidence intervals and regression. Chernick [7] discussed the key ideas and applications of the bootstrap to the above named inferences as well as time series. Young [24] provided an introductory overview to bootstrap and related methods, and discussed bootstrapping for both independent and dependent data.

Banks [5] presented smoothed versions of Efron's bootstrap and Bayesian bootstrap. We provide details of Banks' smoothed version of Efron's bootstrap as this has similarities to NPI-B as introduced in this paper. We call it Banks' bootstrap to clearly distinguish it from Efron's bootstrap. Banks' bootstrap smooths Efron's bootstrap by linear interpolation ('histospline smoothing') between the jump points of the empirical distribution [5]. This procedure for one-dimensional real-valued observations restricted to a finite interval, is as follows. Suppose one has a sample of $n$ observations, for ease of notation and introduction of the methods we assume that there are no tied observations, but any ties are easily dealt with by breaking ties by assuming very small differences or allowing point masses.

1. The $n$ observations create a partition of the finite interval of possible values consisting of $n + 1$ intervals.
2. Select one of these intervals, each with probability $1/(n + 1)$.
3. Sample an observation uniformly from this interval.
4. Repeat steps 2 and 3 $m - 1$ times to create a bootstrap sample of size $m$.

5. Compute the statistic of interest from this bootstrap sample.
6. Repeat steps 2–5 to get multiple bootstrapped values for the statistics of interest; these can be used to derive the bootstrap estimate for the statistic of interest and to quantify variability of this estimate.

Banks [5] provides a detailed study of the performance of his smoothed bootstrap method compared to Efron's bootstrap, and concludes that it tends to perform better, in particular for small data sets.

In this paper, we present an alternative bootstrap method, based on nonparametric predictive inference (NPI). Whilst it shares the smoothing idea with Banks' bootstrap, the procedure differs fundamentally as described later. However, we do not want to restrict its use to a finite interval, and we will discuss in Sect. 2 how to enable the use of NPI-B on the real (half-)line. It should be noted that Banks' bootstrap can similarly be generalized to non-finite support, and this is not considered further in this paper.

It should be emphasized that this paper has two main aims: introducing NPI-B and illustrating its use for NPI for reproducibility of statistical tests. Due to the predictive nature of NPI, and hence of NPI-B, comparison with classic bootstrap methods, which are explicitly aimed at estimation, are complicated; hence we give a brief initial comparison but leave detailed comparison as an important topic for future research. The motivation for NPI-B from our research into test reproducibility is due to the fact that NPI for reproducibility [11] leads to major computational problems for all but small data sets, while explicit expressions for NPI lower and upper reproducibility probabilities can only be derived for a few basic tests. The NPI-B approach provides an attractive solution to these problems, as is explained later in this paper.

Nonparametric predictive inference (NPI) [3, 8, 9] is a frequentist statistical method based on Hill's assumption $A_{(n)}$. Hill [17] introduced the assumption $A_{(n)}$ for prediction if there is no prior information about an underlying population distribution, or, perhaps more realistically, if one prefers not to use any such possible information in order to provide inferences that are strongly based on the data. Assuming that available data consist of $n$ real-valued observations, the assumption $A_{(n)}$ provides direct conditional probabilities for one future real-valued observation, or, when $A_{(n)}, \ldots, A_{(n+m-1)}$ are applied sequentially [1, 9], for $m$ future observations. To introduce $A_(n)$ [17], we denote the $n$ ordered observations by $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$, for ease of notation we define $y_{(0)} = -\infty$ and $y_{(n+1)} = \infty$, which we will replaced by known finite bounds for the possible values of the future observation in case these are known or assumed. These observations partition the real-line into $n+1$ intervals $I_l = (y_{(l-1)}, y_{(l)})$, for $l = 1, 2, \ldots, n+1$. The assumption $A_{(n)}$ provides a partially specified probability distribution for the next observation $Y_{n+1}$ by defining

$$P(Y_{n+1} \in I_l) = \frac{1}{n+1} \tag{1}$$

for $l = 1, 2, \ldots, n+1$.

It is clear that $A_{(n)}$ is a post data assumption related to exchangeability [13], that statistical inferences based on it are predictive and nonparametric, and that it may be suitable if there is no knowledge about the random quantity of interest beyond the data or one explicitly does not wish to use or assume such knowledge. $A_{(n)}$ is not sufficient to get precise probabilities for general events of interest, but it provides lower and upper bounds for a probability for any event and these are lower and upper probabilities in imprecise probability theory [4]. The use of these lower and upper probabilities based on $A_{(n)}$, for a variety of statistical inferences, has been called nonparametric predictive inference (NPI) [3, 8, 9]. Augustin and Coolen [3] presented the NPI lower and upper probabilities for real-valued random quantities, and their strong consistency properties in the theory of imprecise probability. The NPI lower probability for an event $A$ is denoted by $\underline{P}(A)$, the corresponding NPI upper probability is denoted by $\overline{P}(A)$. Generally, $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$. The NPI lower and upper probabilities for the event $Y_{n+1} \in B$, where $B \subset \mathbb{R}$ are

$$\underline{P}(Y_{n+1} \in B) = \frac{1}{n+1} |\{l : I_l \subseteq B\}| \tag{2}$$

$$\overline{P}(Y_{n+1} \in B) = \frac{1}{n+1} |\{l : I_l \cap B \neq \emptyset\}| \tag{3}$$

The lower probability (2) consists of all probability mass, according to $A_{(n)}$, that must be in $B$, while the upper probability (3) consists of all the probability mass that can be in $B$.

Section 2 of this paper presents the NPI bootstrap method together with an initial comparison, via simulation studies, with Efron's bootstrap [14]. The main reason for introducing NPI-B is its application to NPI reproducibility of statistical tests, which is presented in Sect. 3. Test reproducibility is a topic which has received increasing attention in recent years, partly due to some users of statistical methods apparently having difficulties with the interpretation of $p$-values. We have presented inference on test reproducibility from NPI perspective [11], claiming that the predictive nature of NPI fits well with the practical question of interest, namely whether or not a repeat of an experiment would lead to the same overall test result. The exact NPI methods presented in [11] are computationally very demanding for realistic data sets and only applicable for the most basic tests. The NPI bootstrap method presented in this paper provides a suitable tool to implement the NPI approach to test reproducibility to a wider range of tests and larger sample sizes. Sect. 4 contains some concluding remarks.

## 2 NPI Bootstrap

In this section, we present the main idea of NPI bootstrap (NPI-B) for real-valued data, and we provide a brief initial comparison with Efron's bootstrap, mainly to illustrate the differences between the approaches. Detailed comparison for NPI-B with other bootstrap methods is complicated, due to the explicitly

different natures of the approaches: NPI-B is developed for predictive inference while established bootstrap methods are aimed at estimation of population characteristics. Such detailed comparisons to get a complete picture of the advantages and disadvantages of different methods is of course important, e.g., to see if the increased variability in NPI-B compared to other bootstrap methods may also have benefits for quantification of variability of estimates of population characteristics; this is left as an important topic for future research.

In the NPI-B method, the observations are drawn from the intervals between the original data observations, similar to Banks' bootstrap as outlined in Sect. 1. In NPI-B the first bootstrap value is drawn in exactly the same manner as the first value in Banks' bootstrap procedure. However, subsequent values are drawn differently, the key difference being that any already sampled bootstrap observation, for the same bootstrap sample, is added to the data and hence the observations in a single NPI-B bootstrap sample are not conditionally independent, given the original sample of size $n$, as is the case in Banks' bootstrap. So, the first sampled observation is added to the data set, leading to $n+1$ observations which create a partition consisting of $n+2$ intervals. The second observation is then drawn from these $n+2$ intervals, otherwise following the same procedure as in Banks' bootstrap. This is continued, with each observation drawn from the intervals in the partition created by the $n$ original observations together with all previously drawn observations belonging to the same bootstrap sample. This continues until $m$ observations have been drawn, where $m$ is chosen beforehand, and these $m$ observations form one NPI-B sample (which of course does not include the $n$ original data observations).

The NPI-B algorithm for one-dimensional real-valued data on a finite interval is as follows:

1. The $n$ observations create a partition of the finite interval of possible values consisting of $n+1$ intervals.
2. Select one of these intervals, each with probability $1/(n+1)$.
3. Sample an observation uniformly from this interval.
4. Add this observation to the data; increase $n$ to $n+1$.
5. Repeat steps 2–4 (so now with one more data observation then used to sample the previous value), to get a further future value. Stop this once the bootstrap sample consists of $m$ observations.
6. Compute the statistic of interest from this bootstrap sample.
7. Repeat steps 2–6 to get multiple bootstrapped values for the statistics of interest; these can be used to derive the bootstrap estimate for the statistic of interest and to quantify variability of this estimate.

In this algorithm, we use the uniform distribution to sample an observation from a given interval. This distribution can of course be changed, consideration of this option for specific applications where the uniform distribution may not be the most natural assumption is beyond the scope of this introductory presentation of the NPI-B method and is left as an interesting topic for future research.

For our application of NPI-B to test reproducibility, it is important that we can apply the method to real-valued observations without the restriction to a finite interval of possible values for the future observations. The problem is that one cannot sample an observation uniformly from an open-ended interval, so in step 3 of the above algorithm we must assume a different probability distribution over such an interval in order to sample the future observation. Of course, there are many opportunities to do so, and it may be possible to use some background information or additional aspects of the data to choose specific distributions, but to introduce this method we propose to use the tail of a Normal distribution for general real-valued data, and the tail of an Exponential distribution for non-negative real-valued data, e.g., failure time data.

For the first case, so considering data on the whole real-line, we fit a single Normal distribution to be used for both the intervals $\left(-\infty, x_1\right)$ and $\left(x_n, \infty\right)$, and this is done such that, according to this Normal distribution, both these intervals have probability mass $1/(n+1)$, which corresponds to the $A_{(n)}$ assumption. This is achieved by setting the mean $\mu$ and the standard deviation $\sigma$ of the Normal distribution equal to

$$\mu = \frac{x_1 + x_n}{2} \tag{4}$$

and

$$\sigma = \frac{x_n - \mu}{\Phi^{-1}(\frac{n}{n+1})} \tag{5}$$

For the second case, with data on $[0, \infty)$, we fit an Exponential distribution, specified by the cumulative distribution function $P(Y \leq y) = 1 - e^{-\lambda y}$ for $y \geq 0$, to the final interval $(x_n, \infty)$ by setting the rate parameter equal to

$$\lambda = \frac{\ln(n+1)}{x_n} \tag{6}$$

For both these cases, we keep sampling from the Uniform distribution to derive values in any of the other intervals created by the data, which are of finite length.

To get an initial idea of how NPI-B compares to Efron's bootstrap, which we indicate by Ef-B, we report on some results from a simulation study. More details, including some comparison with Banks' bootstrap, can be found in the PhD thesis of the second-named author [6]. It must be emphasized that NPI-B is fundamentally different to the other bootstrap methods due to its explicit aim at predictive inference, while the other methods have all been developed for estimation of population characteristics and related inferences for the quality of the estimates. Hence, detailed comparison of the methods, in particular to see if NPI-B could also be used for the latter objectives, is an important topic for future research.

The different nature of NPI-B compared to Ef-B is clearly seen when considering confidence intervals, related to estimation of particular population characteristics, and prediction intervals, related to predicting a future observation. First,

we compare bootstrap confidence intervals of Ef-B and NPI-B. There are several methods to define bootstrap confidence intervals [20], to illustrate the substantial difference between bootstrap methods for estimation and for prediction they do not make much difference, we will use the BCa interval [15]. Table 1 shows the coverage proportions of $100(1-2\alpha)$ percent BCa intervals for both Ef-B and NPI-B, with data simulated from a Normal distribution with mean 28 and variance 4 (note that for this comparison the values of these two parameters are irrelevant). We have used different values $n$ for the original sample size, with bootstrap samples of the same size, and different values for $\alpha$. We consider estimation of the mean, variance and third quartile ($q_{75}$). We constructed 1000 bootstrap confidence intervals for each case, and for each of these we used 1000 bootstrap samples per run (so for each simulated data set). In the PhD thesis of the second-named author, more distributions are considered, the results lead to the same conclusions as those presented in Table 1, namely that NPI-B performs substantially worse than Ef-B in terms of coverage of confidence intervals for these population characteristics.

It is not unexpected that NPI-B does not provide confidence intervals with the right coverage, and hence performs worse than EF-B, as it is not developed for estimation of population characteristics, but for prediction of future observations, or summary statistics of these. To illustrate this difference, we briefly consider the predictive performance of both these bootstrap methods. We create similar intervals based on the bootstrap methods as the confidence intervals above, but we now compare these with related statistics of a further sample drawn from the assumed distribution, which serves as a future observation of a sample statistic and hence is used to see if it is in the NPI-B or Ef-B prediction intervals.

Again we sample from the Normal distribution as described above, results for other distributions were quite similar [6]. Mojirsheibani [21] and Mojirsheibani and Tibshirani [22] presented different types of bootstrap prediction intervals, including the bootstrap percentile method which we use and which is as follows:

**Table 1** Coverage of $(1-2\alpha)$ confidence intervals for NPI-B and Ef-B

| $n=$ | $\alpha=0.01$ | | | | | $\alpha=0.05$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20 | 50 | 100 | 200 | 500 | 20 | 50 | 100 | 200 | 500 |
| Mean | | | | | | | | | | |
| NPI-B | 0.97 | 0.98 | 0.97 | 0.95 | 0.95 | 0.69 | 0.70 | 0.70 | 0.68 | 0.70 |
| Ef-B | 0.97 | 0.98 | 0.98 | 0.97 | 0.99 | 0.87 | 0.90 | 0.90 | 0.89 | 0.91 |
| Variance | | | | | | | | | | |
| NPI-B | 0.92 | 0.90 | 0.93 | 0.92 | 0.95 | 0.59 | 0.55 | 0.56 | 0.58 | 0.62 |
| Ef-B | 0.91 | 0.96 | 0.97 | 0.98 | 0.98 | 0.83 | 0.87 | 0.88 | 0.90 | 0.91 |
| $q_{75}$ | | | | | | | | | | |
| NPI-B | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 | 0.73 | 0.71 | 0.71 | 0.69 | 0.71 |
| Ef-B | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 | 0.89 | 0.87 | 0.90 | 0.90 | 0.92 |

**Table 2** Coverage of 90% prediction intervals for NPI-B and Ef-B

| $n =$ | 20 | 50 | 100 | 200 | 500 |
|---|---|---|---|---|---|
| Mean | | | | | |
| NPI-B | 0.93 | 0.87 | 0.82 | 0.91 | 0.92 |
| Ef-B | 0.77 | 0.70 | 0.68 | 0.75 | 0.80 |
| Variance | | | | | |
| NPI-B | 0.92 | 0.90 | 0.90 | 0.90 | 0.87 |
| Ef-B | 0.75 | 0.66 | 0.68 | 0.71 | 0.69 |
| $q_{75}$ | | | | | |
| NPI-B | 0.94 | 0.89 | 0.86 | 0.90 | 0.85 |
| Ef-B | 0.80 | 0.71 | 0.70 | 0.77 | 0.74 |

**Table 3** Coverage of 98% prediction intervals for NPI-B and Ef-B

| $n =$ | 20 | 50 | 100 | 200 | 500 |
|---|---|---|---|---|---|
| Mean | | | | | |
| NPI-B | 1.00 | 0.95 | 0.97 | 0.99 | 0.99 |
| Ef-B | 0.92 | 0.88 | 0.89 | 0.93 | 0.93 |
| Variance | | | | | |
| NPI-B | 0.98 | 1.00 | 0.99 | 0.99 | 1.00 |
| Ef-B | 0.78 | 0.89 | 0.91 | 0.90 | 0.88 |
| $q_{75}$ | | | | | |
| NPI-B | 1.00 | 0.95 | 0.99 | 0.98 | 0.99 |
| Ef-B | 0.94 | 0.83 | 0.87 | 0.90 | 0.92 |

1. Draw a sample of size $n$ from a specific distribution, denoted by $x_1, \ldots, x_n$. Then draw a second sample, also of size $n$, from the same distribution, denoted by $y_1, \ldots, y_m$. Let $t_y$ denote the $y$-sample based summary statistic of interest.
2. Use the $x$-sample to draw $B$ bootstrap samples of size $n$ as described above. Calculate the same summary statistic $t_j$ for each of these bootstrap samples, so for $j = 1, \ldots, B$.
3. Construct an $100(1-2\alpha)\%$ prediction interval for $t_y$ by defining the lower bound to be the $\alpha \times B$-th value in the ordered list of the values $t_j$ and the upper bound to be the $(1-\alpha) \times B$-th value in this list (using the nearest integer if these values or not integer).
4. Check if the prediction interval from step 3 contains the value $t_y$ from step 1.

Results for some different cases are given in Tables 2 and 3. These show that the prediction intervals have far better coverage for NPI-B than for Ef-B. It should be emphasized that this comparison does not provide more insight into the performances of these methods beyond this brief initial comparison, which however was supported by more similarly performed comparisons [6]. The difference in performance of NPI-B and Ef-B for estimation and prediction is expected to be

such that NPI-B is better for prediction while Ef-B is better for most estimation scenarios. This is due to the fact that the NPI-B bootstrap samples have far more variability than the Ef-B bootstrap samples, with the former going outside the original data set and positive dependence of values within a single bootstrap sample. Compared to Ef-B, this tends to lead to wider intervals with more variation in the centers of the intervals. Ef-B is well known to work well for estimation of most population characteristics, but it can go wrong if interest is in a very large (or small) percentile of the population distribution, which will typically not be covered well unless the bootstrap samples are large. Ef-B does not cater for the additional variation when considering prediction, for which NPI-B is explicitly developed. NPI-B is expected to perform better when interest is in very large (or small) percentiles, as a reasonable proportion of the bootstrap samples will frequently cover these due to the out-of-data sampling. An important topic for further research into performance of NPI-B, in comparison to Ef-B, is how this depends on the underlying population distribution, in particular skewness might affect both these methods. It is expected that, unless there is extreme skewness, NPI-B keeps performing best for prediction and Ef-B for estimation. A substantially more detailed study into NPI-B, in particular with the possibility to use it, or adapt it, for estimation, is left as an important topic for future research. In this section, we have introduced NPI-B and briefly illustrated its predictive nature, as opposed to Ef-B which is developed for estimation. This was done for the use of NPI-B as a means of advancing the NPI approach for test reproducibility, which is presented in the next section.

## 3 Test Reproducibility Using NPI-Bootstrap

We introduced the NPI method for reproducibility of statistical hypothesis tests in [11]. This has been followed by further investigations, in particular considering tests based on order statistics [10] and likelihood ratio tests [19]. This application of NPI considers the question if a repeat of a statistical hypothesis test, performed in exactly the same way as the actual test, would lead to the same conclusion, that is rejection of the null hypothesis or not. There has been much confusion about test reproducibility, following a paper by Goodman [16] in which the issue was raised and a subsequent discussion to Goodman's paper by Senn [23] in which clarifications from statistical perspectives were provided. For a more detailed introduction to the topic and related literature see [2, 6, 11]. We consider the test reproducibility problem as fundamentally predictive, and hence we have proposed the NPI approach as providing a natural solution to it. The tests considered in our introductory paper [6] were basic nonparametric one-sample tests, while we also considered the two-sample rank sum test (also known as Wilcoxon Mann Whitney test or variations to this name). For the latter, we could only compute NPI lower and upper reproducibility probabilities for small sample sizes, due to the combinatorics involved. Beyond basic tests, which typically have a simple sufficient statistic, it is difficult to derive closed form expressions for the NPI lower

and upper reproducibility probabilities, this is e.g., the case for the two-sample Kolmogorov–Smirnov (KS) test. The NPI Bootstrap method, presented in this paper, is useful for such cases, as we present in this section with specific attention to the KS test. We should note that the PhD thesis of the second-named author [6] also presents the application of NPI-B to the two-sample rank sum test, and compares it for small samples to the analytically derived NPI lower and upper reproducibility probabilities. The NPI-B results are always within the interval created by the NPI lower and upper probabilities, due to the construction of the latter, with no assumptions of probability masses assigned to intervals between consecutive observations, this is a logical result that one would always expect but which in extremely rare cases may not hold, due to the randomness of the bootstrap inferences.

The two-sample Kolmogorov–Smirnov test (KS test) [18] is a well-known nonparametric test for equality of the underlying population distributions of two samples. Suppose that an iid sample $X_1, X_2, \ldots, X_{n_x}$ of size $n_x$ is randomly selected from a population with cumulative distribution function $F_x$, and an iid sample $Y_1, Y_2, \ldots, Y_{n_y}$ of size $n_y$ is randomly selected from a population with cumulative distribution function $F_y$. Consider the null hypothesis $H_0 : F_x(t) = F_y(t)$ for every $t$, versus the alternative hypothesis $H_1 : F_x(t) \neq F_y(t)$ for at least one $t$. Let $\hat{F}_x(t)$ and $\hat{F}_y(t)$ be the empirical cumulative distribution functions based on the $X$ and $Y$ samples, respectively. Let $d$ be the greatest common divisor of $n_x$ and $n_y$ and set $J = \frac{n_x n_y}{d} \max |\hat{F}_x(t) - \hat{F}_y(t)|$, then $J$ is the two-sided two-sample Kolmogorov–Smirnov statistic. To compute it let $H_{(1)}, H_{(2)}, \ldots, H_{(N)}$ be the $N = n_x + n_y$ ordered values of the combined samples $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$. Then $J = \frac{n_x n_y}{d} \max |\hat{F}_x(H_{(i)}) - \hat{F}_y(H_{(i)})|$. At significance level $\alpha$, $H_0$ is rejected if and only if $J \geq j_\alpha$ where $j_\alpha$ can be found in tables [18].

We briefly illustrate the application of NPI-B to the two-sided KS test. Table 4 presents some results for two samples, with $n_x = n_y = 10$, both sampled from the standard Normal distribution, and presenting the NPI-B estimates of the NPI reproducibility probability for 30 simulated cases. For this illustration we use $\alpha = 0.1678$, which leads to test rule that $H_0$ is rejected if and only if $J \geq 5$. This value for the significance level was chosen as it leads to quite similar numbers of cases where the null hypothesis is rejected or not, and hence it provides some meaningful output for the application of our method. As expected, and also seen in earlier NPI studies of test reproducibility [10, 11, 19], the estimates of the NPI

**Table 4** KS test, $H_0$ holds

| $J$ | Frequency | Values of NPI-B-RP |
|---|---|---|
| 2 | 5 | 0.732, 0.734, 0.763, 0.773, 0.781 |
| 3 | 4 | 0.687, 0.689, 0.720, 0.741 |
| 4 | 6 | 0.600, 0.620, 0.624, 0.630, 0.671, 0.693 |
| 5 | 6 | 0.463, 0.463, 0.472, 0.515, 0.519, 0.553 |
| 6 | 5 | 0.518, 0.530, 0.541, 0.589, 0.657 |
| 7 | 4 | 0.674, 0.735, 0.770, 0.774 |

**Table 5** KS test, $H_0$ does not hold

| $J$ | Frequency | Values of NPI-B-RP |
|---|---|---|
| 3 | 3 | 0.611, 0.666, 0.712 |
| 4 | 5 | 0.409, 0.409, 0.451, 0.507, 0.555 |
| 5 | 5 | 0.494, 0.528, 0.607, 0.624, 0.648 |
| 6 | 5 | 0.637, 0.670, 0.718, 0.761, 0.763 |
| 7 | 4 | 0.766, 0.794, 0.803, 0.815 |
| 8 | 6 | 0.897, 0.899, 0.903, 0.906, 0.924, 0.945 |
| 9 | 1 | 0.965 |
| 10 | 1 | 0.980 |

reproducibility probabilities tend to smallest if the original test leads to a test statistic close to the test threshold, so here to a value 4 or 5 for $J$.

Table 5 shows the results of a similar study and the same sample sizes, but now with the $X$ sample drawn from the Uniform distribution on (0, 1) and the $Y$ sample drawn from the Uniform distribution on (0.25, 0.5). Now, as expected, $H_0$ is more often rejected on the basis of the original samples, and we see again that the NPI reproducibility estimates probabilities tend to be larger the further away the original value $J$ is from the test threshold. We also note that, for each value of $J$ in these cases, the entries per table are quite similar; variation is due to the original data samples varying in each case, and of course there is some variation due to the use of the NPI-B procedure. Some further investigations have been reported in the second author's PhD thesis [6], but more detailed investigations are left as important topics for future research. It should be emphasized that the application of NPI-B for reproducibility inference for the KS test was necessary as no closed-form expressions could be derived

## 4 Concluding Remarks

This paper has introduced a new version of bootstrap, nonparametric predictive inference bootstrap (NPI-B), which is explicitly aimed at predictive inference. A brief initial comparison with Efron's bootstrap shows that the latter performs better for estimation but NPI-B performs better for prediction. These initial results motivate much further research, in particular to investigate further properties and performance of NPI-B for other inferences, and also to see if NPI-B can be adapted for use in estimation inferences. A main reason for introducing NPI-B is to enable estimation of test reproducibility probabilities in the NPI framework. While this paper sets out the idea and briefly illustrated it, further investigations on the properties of the method and applications to other scenarios are left as important topics for future research.

**Acknowledgements** The authors thank a reviewer for pointing out important topics for the future investigations related to this work, which have also led to some more clarifications in this paper.

## Compliance with Ethical Standards

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.



## References

1. Arts GRJ, Coolen FPA, van der Laan P (2004) Nonparametric predictive inference in statistical process control. Qual Technol Quant Manag 1:201–216
2. Atmanspacher H, Maasen S (eds) (2016) Reproducibility: principles, problems, practices, and prospects. Wiley, Hoboken
3. Augustin T, Coolen FPA (2004) Nonparametric predictive inference and interval probability. J Stat Plan Inference 124:251–272
4. Augustin T, Coolen FPA, de Cooman G, Troffaes MCM (eds) (2014) Introduction to imprecise probabilities. Wiley, Hoboken
5. Banks DL (1988) Histospline smoothing the Bayesian bootstrap. Biometrika 4:673–684
6. Binhimd S (2014) Nonparametric predictive methods for bootstrap and test reproducibility. PhD Thesis, Durham University, UK. Available from www.npi-statistics.com
7. Chernick MR (2008) Bootstrap methods: a guide for practitioners and researchers. Wiley, Hoboken
8. Coolen FPA (2006) On nonparametric predictive inference and objective Bayesianism. J Log Lang Inf 15:21–47
9. Coolen FPA (2011) Nonparametric predictive inference. In: Lovric M (ed) International encyclopedia of statistical science. Springer, Hoboken, pp 968–970
10. Coolen FPA, Alqifari HN (2018) Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. REVSTAT Stat J 16:167–185
11. Coolen FPA, Binhimd S (2014) Nonparametric predictive inference for reproducibility of basic nonparametric tests. J Stat Theory Pract 8:591–618
12. Davison AC, Hinkley DV (1997) Bootstrap methods and their applications. Cambridge University Press, Cambridge
13. De Finetti B (1974) Theory of probability. Wiley, Hoboken
14. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
15. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, London
16. Goodman SN (1992) A comment on replication, p-values and evidence. Stat Med 11:875–879
17. Hill BM (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. J Am Stat Assoc 63:677–691
18. Hollander M, Wolfe DA (1999) Nonparametric statistical methods. Wiley, Hoboken
19. Marques FJ, Coolen FPA, Coolen-Maturi T (2019) Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests. J Stat Theory Pract 13:15
20. Martin MA (1990) On bootstrap iteration for coverage correction in confidence intervals. J Am Stat Assoc 85:1105–1118
21. Mojirsheibani M (1998) Iterated bootstrap prediction intervals. Stat Sin 8:489–504
22. Mojirsheibani M, Tibshirani R (1996) Some results on bootstrap prediction intervals. Can J Stat 24:549–568
23. Senn S (2002) Comment on 'A comment on replication, p-value and evidence', by S.N. Goodman (Letter to the editor). Stat Med 21:2437–2444 With reply by S.N. Goodman, 2445–2447

24. Young GA (1994) Bootstrap: More than a stab in the dark? (with discussion). Stat Sci 9:382–415